A Predictive Analytics Primer

by Thomas H. Davenport

September 02, 2014

No one has the ability to capture and analyze data from the future. However, there is a way to predict the future using data from the past. It's called predictive analytics, and organizations do it every day.

Has your company, for example, developed a customer lifetime value (CLTV) measure? That's using predictive analytics to determine how much a customer will buy from the company over time. Do you have a "next best offer" or product recommendation capability? That's an analytical prediction of the product or service that your customer is most likely to buy next. Have you made a forecast of next quarter's sales? Used digital marketing models to determine what ad to place on what publisher's site? All of these are forms of predictive analytics.

Predictive analytics are gaining in popularity, but what do you—a manager, not an analyst —really need to know in order to interpret results and make better decisions? How do your data scientists do what they do? By understanding a few basics, you will feel more comfortable working with and communicating with others in your organization about the results and recommendations from predictive analytics. The quantitative analysis isn't magic—but it is normally done with a lot of past data, a little statistical wizardry, and some important assumptions. Let's talk about each of these.

The Data: Lack of good data is the most common barrier to organizations seeking to employ predictive analytics. To make predictions about what customers will buy in the future, for example, you need to have good data on who they are buying (which may

require a loyalty program, or at least a lot of analysis of their credit cards), what they have bought in the past, the attributes of those products (attribute-based predictions are often more accurate than the "people who buy this also buy this" type of model), and perhaps some demographic attributes of the customer (age, gender, residential location, socioeconomic status, etc.). If you have multiple channels or customer touchpoints, you need to make sure that they capture data on customer purchases in the same way your previous channels did.

All in all, it's a fairly tough job to create a single customer data warehouse with unique customer IDs on everyone, and all past purchases customers have made through all channels. If you've already done that, you've got an incredible asset for predictive customer analytics.

The Statistics: Regression analysis in its various forms is the primary tool that organizations use for predictive analytics. It works like this in general: An analyst hypothesizes that a set of independent variables (say, gender, income, visits to a website) are statistically correlated with the purchase of a product for a sample of customers. The analyst performs a regression analysis to see just how correlated each variable is; this usually requires some iteration to find the right combination of variables and the best model. Let's say that the analyst succeeds and finds that each variable in the model is important in explaining the product purchase, and together the variables explain a lot of variation in the product's sales. Using that regression equation, the analyst can then use the regression coefficients—the degree to which each variable affects the purchase behavior—to create a score predicting the likelihood of the purchase.

Voila! You have created a predictive model for other customers who weren't in the sample. All you have to do is compute their score, and offer the product to them if their score exceeds a certain level. It's quite likely that the high scoring customers will want to buy the product—assuming the analyst did the statistical work well and that the data were of good quality.

The Assumptions: That brings us to the other key factor in any predictive model—the assumptions that underlie it. Every model has them, and it's important to know what they are and monitor whether they are still true. The big assumption in predictive analytics is that the future will continue to be like the past. As Charles Duhigg describes in his book *The Power of Habit*, people establish strong patterns of behavior that they usually keep up over time. Sometimes, however, they change those behaviors, and the models that were used to predict them may no longer be valid.

What makes assumptions invalid? The most common reason is time. If your model was created several years ago, it may no longer accurately predict current behavior. The greater the elapsed time, the more likely customer behavior has changed. Some Netflix predictive models, for example, that were created on early Internet users had to be retired because later Internet users were substantially different. The pioneers were more technically-focused and relatively young; later users were essentially everyone.

Another reason a predictive model's assumptions may no longer be valid is if the analyst didn't include a key variable in the model, and that variable has changed substantially over time. The great—and scary—example here is the financial crisis of 2008-9, caused largely by invalid models predicting how likely mortgage customers were to repay their loans. The models didn't include the possibility that housing prices might stop rising, and even that they might fall. When they did start falling, it turned out that the models became poor predictors of mortgage repayment. In essence, the fact that housing prices would always rise was a hidden assumption in the models.

Since faulty or obsolete assumptions can clearly bring down whole banks and even (nearly!) whole economies, it's pretty important that they be carefully examined. Managers should always ask analysts what the key assumptions are, and what would have to happen for them to no longer be valid. And both managers and analysts should continually monitor the world to see if key factors involved in assumptions might have changed over time.

With these fundamentals in mind, here are a few good questions to ask your analysts:

- Can you tell me something about the source of data you used in your analysis?
- Are you sure the sample data are representative of the population?
- Are there any outliers in your data distribution? How did they affect the results?
- What assumptions are behind your analysis?
- Are there any conditions that would make your assumptions invalid?

Even with those cautions, it's still pretty amazing that we can use analytics to predict the future. All we have to do is gather the right data, do the right type of statistical model, and be careful of our assumptions. Analytical predictions may be harder to generate than those by the late-night television soothsayer Carnac the Magnificent, but they are usually considerably more accurate.

Thomas H. Davenport is the President's Distinguished Professor in Management and Information Technology at Babson College, a research fellow at the MIT Initiative on the Digital Economy, and a senior adviser at Deloitte Analytics. He is the author of over a dozen management books, most recently *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines* and *The AI Advantage.*

This article is about ANALYTICS

⊕ Follow This Topic

Related Topics: Data | Decision Making | Knowledge Management

Comments

Leave a Comment

Post Comment

dicksonlam   6 years ago

I do think there is oversimplification of the statistical aspect of predictive analytics. While regression analysis is commonly used, there exists another class of methods that deserve proper mentions. E.g. Bayes Network, Artificial Neural Net, Decision Tree, Support Vector Machine, etc. More importantly, the non-linear analysis aspect and the probability based approach that underpin many of the aforementioned methods.

↱ Reply   2 👍 0 👎

∨ Join The Conversation

POSTING GUIDELINES

We hope the conversations that take place on HBR.org will be energetic, constructive, and thought-provoking. To comment, readers must sign in or register. And to ensure the quality of the discussion, our moderating team will review all comments and may edit them for clarity, length, and relevance. Comments that are overly promotional, mean-spirited, or off-topic may be deleted per the moderators' judgment. All postings become the property of Harvard Business Publishing.